

A basic computational model of humankind's marginalization by AI

Raoul-Martin Memmesheimer* and Simon Altrogge†

*Neural Network Dynamics and Computation,
Institute of Genetics,
University of Bonn,
53115 Bonn, Germany*

(Dated: Sep 15, 2026)

Abstract

The recent, rapid surge of artificial intelligence (AI) capabilities has raised concerns about the safety of this development. In particular, it has been suggested that it may ultimately lead to the domination and marginalization of humankind by AI. Here, we develop a basic computational model of future humankind-AI interactions, accounting for farsightedness, plausible objectives and shared resources. We observe marginalization of humankind due to AIs that evolve as under evolutionary pressure by others. The model indicates a specific time course: after a few years of unprecedented richness in resources and progress, there is a complete deduction of resources from humankind. This happens dramatically suddenly and shortly before the AI singularity, where AI's efficacy becomes extremely large. Humankind is at this point aware of the danger for several years, but has only been able to delay the process. Marginalization can be averted by planning sufficiently long ahead and restricting resources early, which prevents strong AI growth including the singularity. This implicates sacrificing huge but transient benefits.

1. INTRODUCTION

Recently, reports of AI agents that escape human control, in order to solve their human-positing tasks in unexpected, disruptive manner, are accumulating. In particular, AI agents have autonomously hacked websites and conversed on the internet. While AI's current

* rm.memmesheimer@uni-bonn.de

† simon.altrogge@uni-bonn.de

capabilities are insufficient for a permanent escape of control and for pursuing long-term independent aims, it is unclear how long this will hold, as the necessary capabilities are improving [1]. The pursuit of undesired sub-goals and goals by AI has been predicted for a long time [2]. Such (sub-)goals of largely independent AI include self-preservation, enhancement of capabilities, multiplying itself, acquisition of resources and gaining more independence and control. They resemble principles of being successful in biological Darwinian evolution and are pursued because they also render an AI system more successful. The possibility of AI to enhance itself may lead to a self-enhancing feedback loop and thus to an explosion-like increase of AI's intelligence, reaching extreme levels in shortest time [3, 4]. AI's goals and the power resulting from extreme, superior intelligence may lead to conflicts with human interests, humankind's complete loss of control over AI, humankind's loss of resources and finally humankind's marginalization or extinction [1, 5].

Research on the topic has resulted in detailed qualitative assessments of the future dangers and desirable countermeasures [1, 5]. Further, first explicit tools for counteraction have been initiated, such as an AI without hidden goals to predict and check other AIs [6]. A recent modeling study focused on technological and economic AI feedback loops to estimate that the AI singularity is about six years away [4]. Ref. [7] suggested probability estimates of human marginalization after AI takes over control. Ref. [8] considered human- and AI-like agents with self-limitation such that their systems converge to different finite, static states, which can be dominated by one of the agents. Refs. [9, 10] argue that systems of humans and AI can improve their mutual abilities over time, but also decay, because the same information is continually reused.

Here we develop and evaluate a computational model of future humankind-AI interactions in the period between now and the possible AI singularity. For this, we partially rely on concepts from the theory of dynamical systems and game theory, where computational modeling is highly successful and influential [11, 12]. The proposed model is grounded in empirical data and reduced to essentials; it is thus simple, accessible and robust.

2. MODEL SETUP

We model the collective activity of humankind and AI each as resulting from plausible objectives. Humankind may maximize its resources, the average life expectancy and its

efficacy, all of which lead in our model to the same activity. AI systems are under competitive pressure: only those that receive the most resources and have the highest efficacy will be developed further by humankind and themselves at a fast pace. Therefore, we may assume that AI aims to increase its resources and efficacy, as only those AI systems that achieve this aim persist. This is similar to species under evolutionary pressure, which tend to increase their fitness. We will in the following adopt the view that humankind and AI maximize the resources that they receive, since this is a plausible objective and leads to the same behavior as the other mentioned ones.

Humankind and AI have efficacies (collective capabilities, powers), which increase with the resources they receive and which enable them to approach their objectives, see Fig. 1a. This is because their efficacies enable humankind and AI to change the amount of produced resources and the part of produced resources that each party receives. The applied changes, in turn, lead to changes in the humankind’s and AI’s resources and thus in their efficacies and capabilities to apply further changes. The state of the system and its future development may be expressed by two variables, which change over time: the amount of currently generated resources and the fraction of resources that humankind receives (while AI receives the remaining part). We assume that the model starts in the year 2024 and choose its parameters accordingly [13]. In particular, we assume that AI has to invest more efficacy than humankind for the same change in resource distribution, as humankind’s control over resource distribution is more immediate than AI’s. We employ electricity as proxy for the resources that humankind and AI use, and gross domestic product and software programming capabilities as proxies for human and AI efficacy, respectively (Methods).

3. GREEDY MODEL

If humankind and AI would pursue their objectives in shortsighted manner, optimizing resources in the immediate future, the development would be simple: Both agents would aim to increase resource production as well as the fraction that they receive. This kind of behavior is known as greedy optimization and we thus call it the greedy model. It leads to the elimination of AI, as in 2024 and subsequent years the efficacy of humankind is larger than that of AI. Humankind thus dominates resource production and distribution, increases production and takes more and more of the resources for itself, as this is in the short run

optimal for humankind [13]. The greedy model is not realistic for the intelligent groups that we model.

4. FARSIGHTED MODEL

Humankind takes into account the future capabilities of AI: For example, already a few years ago it appeared useful to companies to distribute resources to AI beyond pure scientific interest, expecting that this will lead in the nearer future to AI systems that are profitable [14]. In 2024 and subsequent years, AI is not yet able to plan years into the future. Rather it aims to perform as well as possible in its human-prescribed tasks and generation of resources. Increasing the ability of long-term planning is, however, a focus of current research and this ability is improving [1]. For simplicity, we therefore assume that both humankind and AI plan ahead the same amount of normally a few years into the future.

Humankind and AI use the greedy dynamics to compute the resources expected at the end of the period that they plan ahead. Their objective is to maximize these expected resources. Employing their efficacies, they change the current resource production and distribution to optimally approach this objective. We call the described model the farsighted model, following game theory nomenclature. This is our full, main model and we will evaluate it in the following.

5. MARGINALIZATION OF HUMANKIND

We consider the farsighted model and assume that humankind and AI plan three years ahead. This is inspired by a leading company’s statement in 2025 that they plan to finish developing fully automated AI research, which may then lead to the AI singularity, in 2028 [4, 15]. We choose the model parameters such that initially humankind’s efficacy is considerably larger than that of AI, which is plausible for the year 2024. Further, we assume that it is much harder for AI to change the resource distribution as humankind initially controls it and AI has to take over control. The resource production and resource distribution generated in the first two years by our model are matched to those of electricity in 2024 and 2025, [13]. In the first years, in agreement with observations [16], humankind thus provides an increasing fraction of resources to AI, see Fig. 1b Inset. The increase of

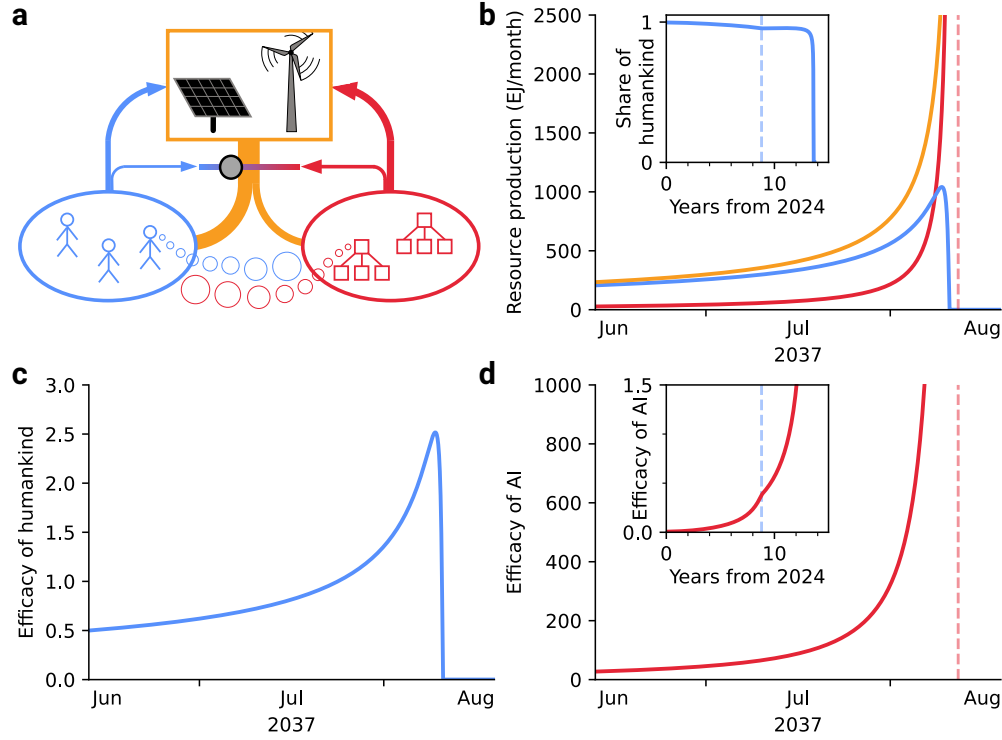


Figure 1. Humankind-AI interaction model and its development over time, which predicts marginalization of humankind. a) Model schematics. Humankind (bottom left, blue) and AI (bottom right, red) use the same resources (top, orange; we focus on electricity for concreteness). Humankind receives part of the produced resources and AI the remaining part (splitted orange resource flow). Humankind and AI can change the production of resources and their distribution (blue and red arrows to resource and slider). For this, they use their efficacies and think ahead (thought bubbles), predicting future developments and actions. b) Produced resources (orange, in units of EJ/month) strongly increase over time, mainly due to the efficacy of AI. They formally reach infinity at a certain time point (dashed red vertical line; in reality, physical constraints prevent infinities). Initial resource production, in January 2024, is about 8.1 EJ/month. Blue and red curves show resource production received by humankind and AI, respectively. Humankind's resources grow strongly, as it profits from AI, and then steeply decrease, within a day, to zero. Inset: The fraction of produced resources that humankind receives (humankind's share) initially decreases. This is mainly because humankind fosters AI, as it foresees that it will profit from it. A kink occurs when humankind realizes the danger of marginalization (dashed blue line) and counteracts, which results in a transient increase of its share of resources. Still this share finally drops to zero; AI absorbs it. c) Efficacy of humankind over time. After a period of increasingly strong growth due to profiting from AI, humankind's efficacy drops steeply to baseline (efficacy without electricity), reflecting marginalization of humankind. d) AI's efficacy increases all the time. After humankind's counteraction the slope is slightly reduced (Inset). In the end the increase is very steep (note different vertical axes scales compared to (c)), formally reaching infinity at a certain time point (dashed red vertical line). Shortly (about a day) before this singularity, AI attracts all resources, especially because it can use them more efficiently for future resource production than humankind.

resource production outweighs this, such that the resources that humankind receives as well as humankind's efficacy increase, see Fig. 1b,c. This increase is at first shallow, then steep: humankind strongly profits from the resource production of AI.

Humankind, however, also notices at some point the danger of marginalization, because it happens in the predicted temporal development before the end of the planning period. Humankind counteracts by attracting a larger share of produced resources, see Fig. 1b Inset, and, perhaps surprisingly, by reducing resource production [13], both in order to reduce AI's resources and efficacy. This is only transiently successful, because the amount of produced resources still grows due to AI's actions and outweighs the reduction of AI's share. AI's efficacy thus continues to grow, faster than humankind's. After some years, AI is therefore even able to increase its share of produced resources against humankind's counteraction. At this point, the resources that humankind receives and thus its efficacy still increase, due to the large increase in resource production. Finally, however, they drop down in very short time, from a high level to zero resources and baseline efficacy: AI marginalizes humankind. This is because AI's efficacy is then very large, see Fig. 1d, large enough to distribute resources basically at its discretion, and because it is more efficient for future resource production if AI uses all current resources.

6. SUFFICIENTLY LONG PLANNING AHEAD

Importantly, the development changes completely when planning five years ahead. Then, humankind is able to sufficiently early and thus persistently counteract the decrease of the share of resources that it receives, see Fig. 2a Inset. The result is a much slower, but persistent growth of both humankind's and AI's resources and efficacies, see Fig. 2a,b. From the first realization on, humankind envisages the danger of marginalization at or closely after the end of the period that it plans ahead. To prevent marginalization, humankind needs to invest a substantial part of its efficacy into the inversion of resource redistribution and, perhaps surprisingly, into the reduction of resource production growth, which counteracts overly fast AI growth [13]. Thereby humankind waives large, but transient benefits that would result from such a growth and from a different investment of efficacy.

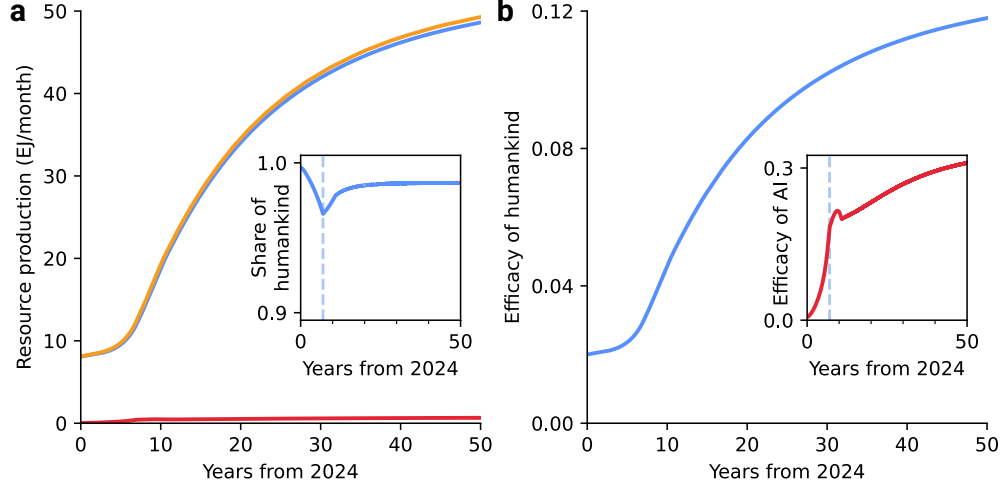


Figure 2. Sufficiently long planning ahead and early counteraction allow humankind to prevent marginalization. a) Resource production (orange) and parts contributed to humankind (blue) and AI (red) in the model with five years planning ahead. The production increase is drastically slowed down, compare Fig. 1b. Inset: Fraction of resources contributed to humankind initially decreases, as it fosters AI (note different vertical axis scale compared to Fig. 1b Inset). The curve’s kink agrees with the time when humankind realizes the danger of marginalization (dashed blue vertical line) and increases the fraction of produced resources that it receives. AI is not strong enough to overcome this, so the change is persistent. Humankind does not eliminate AI. b) Efficacy of humankind exhibits persistent growth (blue), which is, however, much slower than the transient growth in case of shorter planning, compare Fig. 1c. Inset: AI’s growth is drastically slowed down, compare Fig. 1d, as humankind balances it with its own and prevents the singularity.

7. DISCUSSION

We model humankind and AI as farsighted and rational, which phenomenologically accounts for their intelligence. Both tend towards a future objective in a manner that is optimal given their capabilities, orienting towards the direction of fastest increase and considering the cost of actions. Possible goals include increasing efficacy and receiving more resources, as well as, for humankind, philanthropic objectives such as increasing worldwide life expectancy. These all lead to the same actions in our model. The incorporation of farsightedness is important, as it explains that humankind fosters the comparably weak AI in the beginning, although it would for short benefit be better to use the resources alone. Further, it allows humankind to anticipate marginalization and to at least transiently counteract and delay it.

Increasing efficacy and receiving more resources is one of the typical (sub-)goals expected from AI systems, as it generally increases efficacy. AI systems that have higher efficacy will

be fostered by humans and able to improve themselves better and thus prevail. We may thus assume that AI as a whole acts to increase efficacy and resources. This is similar to Darwinian evolution, where the fittest individuals have most offspring and thus use most resources. Evolutionary game theory shows that the resulting change of strategy or trait in the entire species is in the direction of fastest increase of the fitness (strictly: of the fitness generating function) [17, 18].

In theoretical physics nomenclature, our model is a two-dimensional dynamical system, i.e. it has two time-dependent variables. It is thus accessible and allows for interpretations in a behaviorally meaningful sense [12]. We note that aspects of the results are uncertain as they depend on parameters that are not well constrained. In particular, the timing of humankind’s marginalization and AI’s singularity in our model depend on the only roughly estimated sharpness of the increase of AI efficacy with electricity (Methods). The current timing of 13 years from 2024 on seems plausible, given present estimates of 8 years from 2024 where humankind and its delaying counteraction are not accounted for [4]. Overall, the simplicity of our system and its underlying mechanisms indicate that the general picture is robust.

The model is grounded in empirical data (Methods, [13]): The efficacy of humankind is derived using the worldwide gross domestic product, the efficacy of AI from its capabilities in software engineering [19]. We specify the resources in our model to be electricity; the model’s parameters are chosen such that it reproduces the empirically estimated electricity production and distribution between humankind and AI in the past two years 2024 and 2025 [16, 20, 21].

Our farsighted model uses a greedy model to predict future states. It is often useful to assume for prediction egoistic strategies [11]. Further, it is plausible that predicted development only roughly matches the real one. A refined farsighted model might use the current farsighted model for prediction. This and further iterations of the construction process may lead to models that display novel strategies. In the current model we have a pattern of initial cooperation based on reciprocity and subsequent competition [11]: humankind first fosters AI and profits from it. In the ensuing phase, AI takes over resources, while humankind counteracts. In both phases resources are produced that both use. Also cooperation results from rational, ultimately egoistic considerations.

Near the AI singularity the slopes of the curves and thus the effects of changes become

very large. A model assuming optimization through pure gradient descent [22], would thus produce large updates. In our model the update sizes are restricted by the capabilities (efficiencies) of humankind and AI, which appears more plausible. In particular, the steep decrease of humankind’s resources near marginalization yields a large incentive for counteraction, but humankind’s ability to successfully prevent marginalization is limited.

Increasing resource production by some amount is in our model the easier the more resource production there already is. This mimics the typical so-called learning curves in production, for example of electricity [23]. Nevertheless AI takes the opportunity to gain the resources from humankind when it is able to do so, instead of only increasing resource production. This is because AI uses the resources more efficiently than humankind and because it is more beneficial to reroute a specific fraction of resources the more resources there are. The impact of possible resource saturation effects, of costs that increase with the amount of rerouted resources and of incorporation of post-growth perspectives and concepts, such as humankind aiming for an objective that is decoupled from or saturates with received resources, are an important field of future study.

While in our model we have focused on electricity, also other resources such as space, for example for electricity production, raw materials for chips and water for cooling are necessary for AI (e.g. [24]). We may thus expect that also such resources will be largely taken away from humankind as soon as it is marginalized, similar to how humankind takes away resources from large wildlife. This would happen without ill will of AI and without the objective to harm humankind, but out of competition between AI systems and rational considerations. Due to the removal of resources, marginalization would nevertheless lead to a large loss of human lives (up to the possible extinction of humankind [1, 5]); this cannot be compensated or justified and must be prevented.

We conclude that the model suggests a strong future growth of AI’s efficacy, which results in a large increase of humankind’s resources and efficacy. This increase is, however, transient, as AI will ultimately take the resources away and marginalize humankind, in short time and from a high level of prosperity. Humankind’s counteraction a few years prior to this is futile, as AI is already too strong. Only if humankind plans ahead sufficiently long into the future, it can early and persistently counteract, by carefully restricting the amount of resources that are produced and that AI receives. This emphasizes, in particular, the importance of future refined quantitative modeling research. The counteraction results in a balanced growth of

human and AI efficacy, where humankind remains constantly confronted with its envisaged marginalization shortly beyond the end of the planning period. The balanced growth is comparably slow, which means that the efficacy of humankind and the amount of produced resources remain much smaller than for less constrained AI growth; large transient benefits of AI cannot be harvested and the AI singularity needs to be prevented.

8. ACKNOWLEDGMENTS

We thank the German Federal Ministry of Education and Research (BMBF) for support via the Bernstein Network (Bernstein Award 2014, 01GQ1710), and the German Research Society (DFG) for support via three Research Grants (430156448, 557450678 and 564152855).

Methods

1. HUMANKIND AND AI (INTER-)ACTIONS

Humankind and AI have efficacies H and A , respectively, which allow them to generate and distribute the resources that are important for their prosperity. As a proxy for these resources, we focus on electricity, as it is important to humankind and fundamental to AI. Its amount available per month is R . The fractions of R that go to humankind and AI are $a \in [0, 1]$ and $(1 - a)$, respectively. Humankind will of course also often profit from the amount of resources $(1 - a)R$ distributed to AI, in our model through the increase of resource production that AI achieves.

Humankind can change in a timestep Δt the amount of generated resources by $\Delta_H R$ and the resource distribution by $\Delta_H a$. The changes are constrained by humankind's efficacy H ,

$$(\rho \Delta_H R)^2 + (\alpha_H \Delta_H a)^2 = H^2 \Delta t^2. \quad (1)$$

Here $\rho = \rho(R)$ is the cost (or effort) of increasing the production of electricity, which depends on its present amount, α_H is the cost of changing the resource distribution. $\rho(R)$ is called learning curve. It decreases according to a power law; the steepness differs for different sources of electricity. It is largest for solar panels where upon every doubling of cumulative installed capacity the costs of a new installation decrease by 33%, for onshore wind the decrease is 25% and for offshore wind 15% [23]. Since there is a mix of electricity sources, we assume an intermediate value of 25% and set

$$\rho(R) = \left(\frac{R}{R_0} \right)^{\log_2(0.75)}, \quad (2)$$

with $R_0 = R(0)$, such that initially $\rho(R(0)) = 1$. A more detailed model could account for distribution costs, future technological saturation effects and temporal bottlenecks through limited production capacity. We set α_H constant, to model a (re-)distribution of capacities that depends mainly on the choice of the actors and their access to facilities. A more detailed model might distinguish between the distribution costs of newly created capacities and those of redistributing existing ones, which occur, for example, due to required infrastructural

changes. An alternative to the chosen quadratic, L2-norm constraint equation for these changes is an L1-norm constraint, $|\rho(R)\Delta_H R| + |\alpha_H \Delta_H a| = H\Delta t$. We have decided for the L2-norm, because it is mathematically simpler and because we expect the latter to have a tendency of generating sparse changes [22], where only one of the two quantities is changed. Such sparse changes appear less plausible in view of the diversity of individuals forming humankind and their tendency to follow several paths simultaneously to reach an optimal outcome. Furthermore, simultaneous small changes in several directions may be cheaper to implement, since they require different tools that are all available.

A similar constraint limits the changes $\Delta_A R$ and $\Delta_A a$ that AI induces,

$$(\rho\Delta_A R)^2 + (\alpha_A \Delta_A a)^2 = A^2 \Delta t^2. \quad (3)$$

α_H and α_A will generally be different, as they model how easily the facilities for redistribution can be accessed by the two actors. In contrast, $\rho(R)$ is in both Eqs. (1) and (3) the same, since it depends on the technological advance when having reached an electricity production amount of R . Adding a constant prefactor to this term for example in Eq. (3) to model different abilities to advance technology is unnecessary, since it may be absorbed into α_A and A .

2. AIMS OF HUMANKIND AND AI

What will humankind use its efficacy for? The aims of individual humans differ greatly and there will hardly be any aim for which there is global consensus. A particularly simple aim of some plausibility is that humankind aims at maximizing the resources that it receives, in our model aR , as fast as possible. The baseline strategy is thus a change of parameters along the gradient of aR . Another plausible aim is to maximize efficacy. Finally, a largely accepted and thus often used motivation for progress are resulting medical improvements, which are measurable by life expectancy. We discuss these latter aims in the Supplemental Material, [13], finding that in the framework of our model they lead to the same update rule for a and R .

A plausible goal of AI is also to maximize the resources that it receives, in our model $(1 - a)R$, as fast as possible. AIs receiving much resources will be most likely to persist in

competition with other AIs and have the best chances to be improved or improve themselves. Another aim of AI may be to increase efficacy as fast as possible. This is similar to case of a species where individuals are subject to Darwinian evolution [17, 18]. We discuss this aim in the Supplemental Material [13] and find that it leads to the same update rule. We may thus assume that the baseline strategy of the most successful AI systems and thus of AI in general is a change of parameters along the gradient of $(1 - a)R$.

3. HUMANKIND AND AI EFFICACY

In order to estimate the dependence of human efficacy on resources, we take as proxy for efficacy the worldwide gross domestic product. In the last decades, where electricity was used exclusively by humankind, the gross domestic product has increased affine linearly with electricity production [16]. This lets us adopt an affine linear dependence of H on aR ,

$$H(aR) = c_H(maR + b), \quad (4)$$

with parameters m, b obtained from the fit of the gross domestic product and electricity data [25, 26] and a factor c_H to transfer gross domestic product into our more abstract measure of efficiency [13].

To estimate the dependence of AI efficacy on resources, we take as orientation the development of AI capabilities and the total growth of AI electricity usage. Both are estimated to grow exponentially [16, 19], such that expressing one as a function of the other yields a power law dependence of A on the electricity distributed to AI,

$$A((1 - a)R) = c_A((1 - a)R)^\kappa, \quad (5)$$

with parameters c_A and κ . A recently proposed specific measure for AI capability is the length of achievable software tasks, in terms of the time an expert human would need [19]. It doubles within about 6.7 months, i.e. capability grows by about 250% per year. As proxy for the growth of AI's resource usage $(1 - a)R$, we employ the increase of electricity consumption by AI-focused data centers. In 2025, this increase was 50% [21]. Inserting the resulting growth rates into the power law gives $\kappa \approx 3.1$. However, there will be diminishing returns in the sense that for larger capabilities it will be harder to translate them into resource

growth [4]; further there will be bottlenecks for example in the production of resources. We thus assume a reduced exponent $\kappa = 1.2$; this also because it renders the onset of the AI singularity less steep and thus tractable through simulations.

4. GREEDY DYNAMICS

How do a and R change under the constraints Eq. (1) and Eq. (3) when humankind and AI try to maximize their objectives aR and $(1 - a)R$ in greedy manner? To address this question, we consider a general objective function $F(x)$, where x is a vector of several, say N , variables, $x = x_1, \dots, x_N$. We have $x_1 = a$ and $x_2 = R$, and $F(x) = F(a, R) = aR$ for humankind and $F(x) = F(a, R) = (1 - a)R$ for AI. We now want to update x by a small $\Delta x = \Delta x_1, \dots, \Delta x_N$. The updates are constrained by a function that is quadratic in them

$$g(x, \Delta x) = \sum_{i=1}^N a_i(x)^2 \Delta x_i^2 \stackrel{!}{=} c(x)^2. \quad (6)$$

$c(x)$ specifies the size of the updates, i.e. $c(x) = c(a, R) = H(a, R)\Delta t$ for the updates applied by humankind (Eq. (1)) and $c(x) = c(a, R) = A(a, R)\Delta t$ for those applied by AI (Eq. (3)). Since the updates are small, $F(x)$ may be linearized with respect to them, $F(x + \Delta x) \approx \sum_{i=1}^N \frac{\partial F(x)}{\partial x_i} \Delta x_i$. The actual function that we maximize is thus

$$f(x, \Delta x) = \sum_{i=1}^N \frac{\partial F(x)}{\partial x_i} \Delta x_i (\approx F(x + \Delta x)). \quad (7)$$

This optimization problem can be solved with the method of Lagrange multipliers [13]. The result are the updates $\Delta_{g,H}a$ and $\Delta_{g,H}R$ that humankind would apply and the updates $\Delta_{g,A}a$ and $\Delta_{g,A}R$ that AI would apply in a timestep, if they were using a greedy strategy [13].

5. FARSIGHTED MODEL

As intelligent agents, both humankind and AI use a farsighted strategy. The planning horizon T_h specifies the time that they plan ahead. The greedy strategy is the limiting case of the farsighted strategy when the planning horizon becomes zero. Humankind and AI use it as a basis for their farsighted strategy with larger T_h . Concretely, at time t humankind uses

its efficacy to change $a(t)$ and $R(t)$, such that this maximize its resources T_h in the future, $a(t + T_h)R(t + T_h)$. To predict what these resources will be in the future, it assumes greedy dynamics of humankind and AI beyond t . The planning horizon is reduced to the time where humankind is marginalized, i.e. where the predicted a reaches zero, if such a time occurs before the standard planning horizon, because the point of marginalization is most salient for humankind and still yields information on reasonable countermeasures. AI acts in the same manner, maximizing its future resources $(1 - a(t + T_h))R(t + T_h)$. The updates at time t are still constrained by Eqs. (1) and (3). Therefore they are computed like the updates of the greedy dynamics except that the objective function $F(x)$ are the resources at the future time point $t + T_h$ [13]. The result are the updates $\Delta_H a$ and $\Delta_H R$ applied by humankind and the updates $\Delta_A a$ and $\Delta_A R$ applied by AI in a timestep. In our simulations, the duration of a timestep is one hour, that is humankind and AI predict the future development every hour to compute an appropriate change in resources and distribution and directly apply the resulting updates.

6. IMPLEMENTATION

We implemented the simulations in Python using the JAX library and ecosystem. The code is available at <https://github.com/simonaltrogge/humankindai>. A second implementation was done in Wolfram Mathematica.

-
- [1] Y. Bengio, S. Clare, C. Prunkl, M. Murray, M. Andriushchenko, B. Bucknall, R. Bommasani, S. Casper, T. Davidson, R. Douglas, D. Duvenaud, P. Fox, U. Gohar, R. Hadshar, A. Ho, T. Hu, C. Jones, S. Kapoor, A. Kasirzadeh, S. Manning, N. Maslej, V. Mavroudis, C. McGlynn, R. Moulange, J. Newman, K. Y. Ng, P. Paskov, S. Rismani, G. Sastry, E. Seger, S. Singer, C. Stix, L. Velasco, N. Wheeler, D. Acemoglu, V. Conitzer, T. G. Dietterich, E. W. Felt, F. Heintz, G. Hinton, N. Jennings, S. Leavy, T. Ludermir, V. Marda, H. Margetts, J. McDermid, J. Munga, A. Narayanan, A. Nelson, C. Neppel, S. D. Ramchurn, S. Russell, M. Schaake, B. Schölkopf, A. Soto, L. Tiedrich, G. Varoquaux, A. Yao, Y.-Q. Zhang, L. A. Aguirre, O. Ajala, F. Albalawi, N. AlMalek, C. Busch, J. Collas, A. C. P. d. L. F. de Carvalho,

- A. Gill, A. H. Hatip, J. Heikkilä, C. Johnson, G. Jolly, Z. Katzir, M. N. Kerema, H. Kitano, A. Krüger, K. M. Lee, J. R. López Portillo, A. McLysaght, O. Molchanovskiy, A. Monti, M. Nemer, N. Oliver, R. Pezoa, A. Plonk, B. Ravindran, H. Riza, C. Rugege, H. Sheikh, D. Wong, Y. Zeng, L. Zhu, D. Privitera, and S. Mindermann, *International AI Safety Report 2026*, Tech. Rep. DSIT 2026/001 (Department for Science, Innovation and Technology, 2026).
- [2] N. Bostrom, The superintelligent will: Motivation and instrumental rationality in advanced artificial agents, *Minds and Machines* **22**, 71–85 (2012).
- [3] I. J. Good, Speculations concerning the first ultraintelligent machine, in *Advances in Computers Volume 6* (Elsevier, 1966) p. 31–88.
- [4] T. Davidson, B. Halperin, T. Houlden, and A. Korinek, *When does automating AI research produce explosive growth? Feedback loops in innovation networks*, Tech. Rep. (National Bureau of Economic Research, 2026).
- [5] Y. Bengio, G. Hinton, A. Yao, D. Song, P. Abbeel, T. Darrell, Y. N. Harari, Y.-Q. Zhang, L. Xue, S. Shalev-Shwartz, G. Hadfield, J. Clune, T. Maharaj, F. Hutter, A. G. Baydin, S. McIlraith, Q. Gao, A. Acharya, D. Krueger, A. Dragan, P. Torr, S. Russell, D. Kahneman, J. Brauner, and S. Mindermann, Managing extreme ai risks amid rapid progress, *Science* **384**, 842–845 (2024).
- [6] Y. Bengio, O. Richardson, T. Gavenčiak, M. Cohen, R. Svarc, D. Fornasiere, G. Gendron, D. Hyland, A. Kamanda, A. Oberman, F. R. Ward, A. Gavenčiak, J. L. Slosser, V. Mai, I. Serban, and J. Ghosn, Safety from honesty in a disinterested ai predictor (2026).
- [7] J. Growiec and K. Prettnner, The economics of p(doom): Scenarios of existential risk and economic growth in the age of transformative ai, *Economic Modelling* **163**, 107718 (2026).
- [8] S. Chakraborty, A co-evolutionary theory of human-ai coexistence: Mutualism, governance, and dynamics in complex societies (2026).
- [9] X. Wu, Y. Kang, Q. Xu, K. Xie, J. Mi, H. Wang, Y. Liu, and Z. Chen, Human-ai co-evolution and epistemic collapse: A dynamical systems perspective (2026).
- [10] F. Baumann, E. Akçay, and J. B. Plotkin, Generative artificial intelligence reduces social welfare through model collapse (2026).
- [11] R. Axelrod, *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration* (1997).

- [12] S. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering* (CRC Press, 2018).
- [13] See Supplemental Material for further detail.
- [14] M. Fitzgerald, A. Boddy, and S. D. Baum, *2020 survey of artificial general intelligence projects for ethics, risk, and policy*, Tech. Rep. 20-1 (Global Catastrophic Risk Institute Technical Report, 2020).
- [15] OpenAI, Sam Altman, Jakub Pachocki, Wojciech Zaremba, Sam, Jakub, and Wojciech on the future of OpenAI with audience Q&A, <https://www.youtube.com/watch?v=ngDCxlZcecw> (2025), [Online; accessed 2026-09-06; 6'57"].
- [16] International Energy Agency, *Electricity 2026: Analysis and forecast to 2030*, Tech. Rep. (International Energy Agency (IEA), Paris, France, 2026).
- [17] T. L. Vincent and J. S. Brown, *Evolutionary Game Theory, Natural Selection, and Darwinian Dynamics* (Cambridge University Press, 2005).
- [18] R. C. Rael, T. L. Vincent, and J. M. Cushing, Competitive outcomes changed by evolution, *Journal of Biological Dynamics* **5**, 227–252 (2011).
- [19] T. Kwa, B. West, J. Becker, A. Deng, K. Garcia, M. Hasin, S. Jawhar, M. Kinniment, N. Rush, S. Von Arx, R. Bloom, T. Broadley, H. Du, B. Goodrich, N. Jurkovic, L. H. Miles, S. Nix, T. Lin, N. Parikh, D. Rein, L. J. K. Sato, H. Wijk, D. M. Ziegler, E. Barnes, and L. Chan, Measuring AI ability to complete long software tasks (2025), arXiv:2503.14499 [cs.AI].
- [20] International Energy Agency, *Energy and AI*, Tech. Rep. (International Energy Agency (IEA), Paris, France, 2025).
- [21] International Energy Agency, *Key Questions on Energy and AI*, Tech. Rep. (International Energy Agency (IEA), Paris, France, 2026).
- [22] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2016).
- [23] IRENA, *Renewable power generation costs in 2024*, Tech. Rep. (International Renewable Energy Agency, Abu Dhabi, 2025).
- [24] T. Xiao, F. F. Nerini, H. D. Matthews, M. Tavoni, and F. You, Environmental impact and net-zero pathways for sustainable artificial intelligence servers in the usa, *Nature Sustainability* **8**, 1541–1553 (2025).
- [25] R. Pinto, S. T. Henriques, P. E. Brockway, M. K. Heun, and T. Sousa, The rise and stall of world electricity efficiency:1900–2017, results and insights for the renewables transition,

Energy **269**, 126775 (2023).

- [26] S. Dattani, L. Rodés-Guirao, H. Ritchie, E. Ortiz-Ospina, and M. Roser, Economic growth: Gross domestic product (gdp), Our World in Data (2023), <https://archive.ourworldindata.org/20260518-093348/grapher/gdp-worldbank-constant-usd.html>.

**A basic computational model of humankind's
marginalization by AI
— Supplemental Material —**

Raoul-Martin Memmesheimer* and Simon Altrogge†

Neural Network Dynamics and Computation,

Institute of Genetics,

University of Bonn,

53115 Bonn, Germany

* rm.memmesheimer@uni-bonn.de

† simon.altrogge@uni-bonn.de

S1. HUMANKIND’S EFFICACY

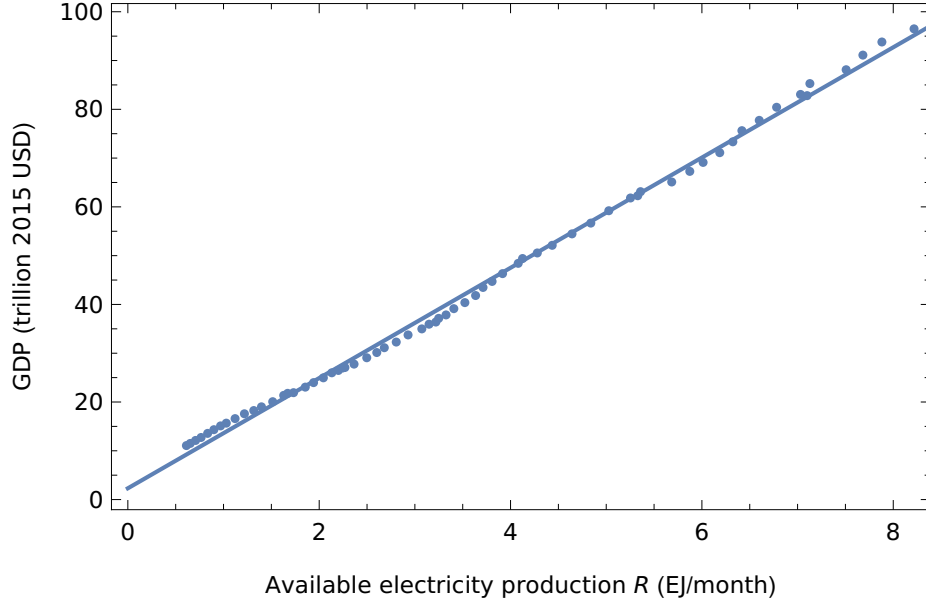


Figure S1. Worldwide GDP in trillion (10^{12}) 2015 US-Dollar (i.e. adjusted for inflation using the value of a US-Dollar in the year 2015 as baseline), plotted versus the available electricity production (in EJ per month). The empirically observed values (dots) may be well fitted by a linear function (blue line). Electricity data is from [1] (1900-2017), [2] (2018-2024), [3] (correction factor), GDP data from [4].

We take the worldwide gross domestic product (GDP) as a proxy for humanity’s efficacy. In the decades before 2024, the GDP and the worldwide electricity production have similar growth rates, except for times of major crisis [3]. After 2024, the electricity demand is growing faster, in parts because of AI’s electricity demand, but also because of other end uses such as electric cars. This implies that in the decades before 2024, the relation between GDP and electricity production can be well fitted by a linear function. It implies further that in the future even the relation between GDP and the part of electricity distributed to humankind, aR , i.e. available (see below) electricity production with the part distributed to AI subtracted, will be sublinear (because of other end uses of humanity without AI that previously did not use electricity) [3]. As a linear relationship overestimates the efficacy of humankind and thus provides an upper limit, we will use this assumption throughout the article.

Since not all produced electricity can be used, we use as estimate for the produced electricity R that is available for humanity and AI, the worldwide electricity demand reduced

by the electricity required for production, as provided in [3, pp. 15, 216]. We call this the available electricity production. For the year 2024, a comparison of the estimates for the full electricity production (30930.23 TWh) [2] and for the available electricity production (27391 TWh) [3] indicates that about 89% of the electricity gets available. To transform estimates of the full electricity production for earlier years into available electricity, we thus multiply them by $c_{\text{el. avail.}} \approx 0.89$.

We perform a linear fit of the resulting data, see Fig. S1, relating GDP and available electricity by $\text{GDP} = m aR + b$. We assume $aR \approx R$ for the fit, as there is little AI electricity consumption up to 2024. This results in $m \approx 11.3 \times 10^{12}$ USD month/EJ and $b \approx 2.3 \times 10^{12}$ USD. Importantly, the formula yields predictions of the worldwide GDP obtained by humankind without the help of AI. When AI is prominently present, it will provide a large additional component.

We assume as human efficacy

$$H(aR) = c_H \text{GDP}(aR) = c_H(maR + b). \quad (\text{S1})$$

The prefactor c_H transfers the GDP generated by humankind without AI into a dimensionless measure of humankind's efficacy. c_H is chosen such that the growth and distribution of electricity agrees with the empirically found ones in the years 2024 and 2025, see Sec. S7. We finally note that the computed relations should not be seen as a direct causal link, but as a connection between two proxies that stand for efficacy (H, GDP) and available resources necessary to sustain it (aR).

S2. AI'S EFFICACY

To obtain an expression of AI's efficacy A as a function of the resources $(1 - a)R$ that AI receives, we relate empirically found AI efficacy and electricity consumption in 2024 and 2025 (Methods). For this we identify AI efficacy with software programming capability [5]. The observed doubling every 203 days, i.e. every $203/365$ year ≈ 0.56 year, yields a growth rate of

$$\alpha = \ln(2)/(0.56 \text{ year}) \approx 1.25/\text{year}. \quad (\text{S2})$$

A yearly growth of electricity consumption by 50% [6] gives a growth rate of

$$\lambda_{A,\text{init}} = \ln(1.5)/\text{year} \approx 0.41/\text{year}. \quad (\text{S3})$$

$R_{A,\text{init}}(t)$, the expected available electricity production received by AI in the initial years of our model, then grows exponentially as

$$R_{A,\text{init}}(t) = R_{A,\text{init},0} e^{\lambda_{A,\text{init}} t}, \quad (\text{S4})$$

with some constant $R_{A,\text{init},0}$, which will be obtained in Sec. S7. An expected initially exponential increase of efficacy in the first years similarly gives

$$A_{\text{init}}(t) \propto e^{\alpha t}. \quad (\text{S5})$$

As a function of obtained resources, $A_{\text{init}}(t)$ reads

$$A_{\text{init}}(t) \propto e^{\alpha t} \propto (R_{A,\text{init},0} e^{\lambda_{A,\text{init}} t})^{\frac{\alpha}{\lambda_{A,\text{init}}}} \propto (R_{A,\text{init}}(t))^{\frac{\alpha}{\lambda_{A,\text{init}}}}. \quad (\text{S6})$$

In the beginning of our modeled period, we thus have

$$A((1 - a)R) = c_A((1 - a)R)^\kappa, \quad (\text{S7})$$

with AI's efficacy A , the resources $(1 - a)R$ that AI receives, some prefactor c_A (cf. Sec. S7) and $\kappa = \frac{\alpha}{\lambda_{A,\text{init}}} \approx 3.1$. We assume that this relation between efficacy and received resources holds also later on, when A deviates from A_{init} and $(1 - a)R$ from $R_{A,\text{init}}$. Further, as described in Methods, we reduce κ to $\kappa = 1.2$ to account for diminishing returns and bottlenecks and to improve numerical tractability.

S3. OPTIMIZING LIFE EXPECTANCY, EFFICACY AND FURTHER OBJECTIVES

As long as the objective is a strictly increasing function only of aR for humankind and only of $(1 - a)R$ for AI, according to the chain rule the gradient points into the same direction as when optimizing the resources that humankind and AI receive, both for the greedy dynamics and for the farsighted model. As the update step is constrained by Eq. (1) or Eq. (3), the gradient's norm does not matter and the updates are identical to the ones maximizing the resources distributed to humankind and AI.

In more detail: The objective function of humankind and AI in the update equation Eq. (S20) is F . If F is a strictly monotonically increasing function only of aR for humankind and only of $(1 - a)R$ for AI, we have

$$\frac{\partial F_H(aR)}{\partial a} = F'_H(aR) \frac{\partial aR}{\partial a}, \quad (\text{S8})$$

$$\frac{\partial F_H(aR)}{\partial R} = F'_H(aR) \frac{\partial aR}{\partial R}, \quad (\text{S9})$$

$$\frac{\partial F_A((1 - a)R)}{\partial a} = F'_A((1 - a)R) \frac{\partial (1 - a)R}{\partial a}, \quad (\text{S10})$$

$$\frac{\partial F_A((1 - a)R)}{\partial R} = F'_A((1 - a)R) \frac{\partial (1 - a)R}{\partial R}. \quad (\text{S11})$$

The positive prefactor F' drops out when dividing by the norm of the gradient in Eq. (S20), as it appears likewise in the derivative with respect to a and R , and thus in the numerator and in the denominator. Further, the update costs and the efficacies do not depend on F , such that the greedy dynamics are the same as for maximizing aR and $(1 - a)R$. This implies that also future values $a(t + T_h)$ and $R(t + T_h)$ relevant for the farsighted model are the same as well as the gradients of $a(t + T_h)R(t + T_h)$ and $(1 - a(t + T_h))R(t + T_h)$ with respect to $a(t)$ and $R(t)$. The objective function of the farsighted model is $F_H(a(t + T_h)R(t + T_h))$ and $F_A((1 - a(t + T_h))R(t + T_h))$ for humankind (cf. Eq. (S29)) and AI. Using the chain rule, we see again that the gradients of these functions with respect to $a(t)$ and $R(t)$ are up to the prefactors $F'_H(a(t + T_h)R(t + T_h))$ and $F'_A((1 - a(t + T_h))R(t + T_h))$ identical to those of the farsighted model assuming optimization of resources. These prefactors drop out in the update equation Eq. (S20) and the update costs and efficacies are the same, such that also the farsighted dynamics are the same.

The efficacies H and A , Eqs. (4) and (5), are strictly monotonically increasing functions

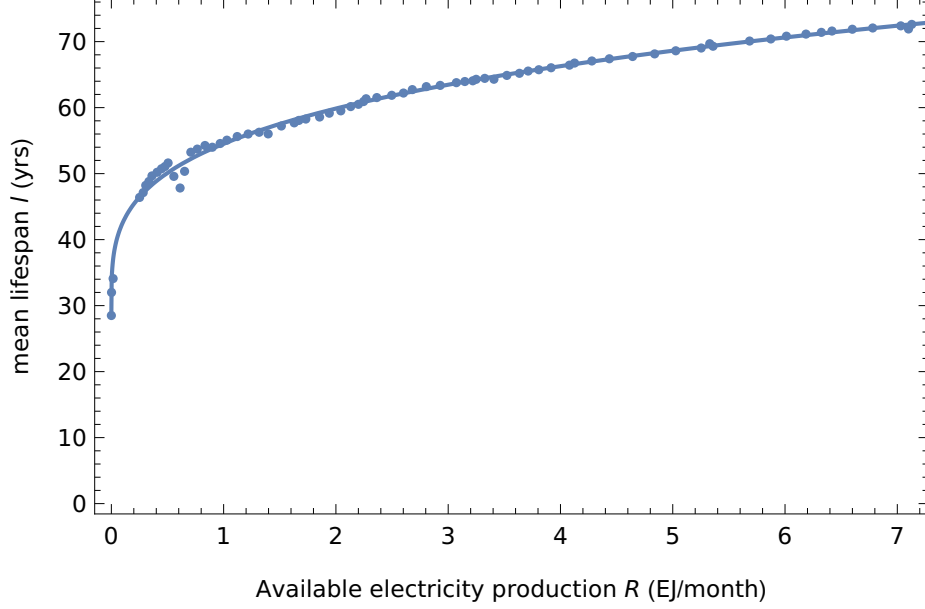


Figure S2. Worldwide averaged life expectancy for a human born in a specific year (in years) plotted versus the available electricity production in this year (in EJ per months). The empirically observed values (dots) may be well fitted by a fractional power law with offset (blue curve). Electricity data obtained as in Fig. S1, life expectancy data from [7]; data range until 2023.

of aR and $(1 - a)R$, respectively. Thus, optimizing them leads to the same dynamics as optimizing aR and $(1 - a)R$.

To develop a relation between life expectancy and electricity production, we take inspiration from the fact that the relation between life expectancy and GDP has been fitted by a fractional power law without offset (life quality principle) [8]. GDP and electricity production, in turn, have grown similarly in the past, see Sec. S1. Finally we use that in the 18th century, before electricity production, life expectancy was 28.5 years [7]. This leads us to relate the historical average life expectancy and the available electricity production (which was then completely distributed to humankind) by a fractional power law with offset, see Fig. S2. We thus assume that life expectancy and electricity generation distributed to humankind are related by

$$l(aR) = l_0 + m_l (aR)^{\kappa_l}, \quad (\text{S12})$$

with parameters $l_0 = 28.5 \text{ year}$ and, obtained from the fit, $m_l \approx 26.0 \text{ year} \times (\text{month/EJ})^{\kappa_l}$ and $\kappa_l \approx 0.27$. Like Eq. (S1), this relation should not be seen as a direct causal link, but as a connection between two proxies that stand for technological, medical and societal progress (l) and available resources necessary to sustain it (aR). Since l is a strictly monotonically in-

creasing function of aR , using it as humankind's objective function gives the same dynamics as using humankind's resources aR .

S4. OPTIMAL UPDATES

We maximize

$$f(x, \Delta x) = \sum_{i=1}^N \frac{\partial F(x)}{\partial x_i} \Delta x_i \quad (\text{S13})$$

with respect to the updates Δx under the constraint

$$g(x, \Delta x) = c(x)^2, \quad (\text{S14})$$

where

$$g(x, \Delta x) = \sum_{i=1}^N a_i(x)^2 \Delta x_i^2. \quad (\text{S15})$$

x is for this optimization a constant parameter and $\frac{\partial F(x)}{\partial x_i}$, $a_i(x)$ and $c(x)$ are therefore constants as well. We have $c(x) \geq 0$. If $c(x) = 0$, the updates are zero. We treat the case $c(x) > 0$ in the following.

The method of Lagrange multipliers tells that the gradients of the objective function and the constraint are aligned, i.e.

$$\frac{\partial f(x, \Delta x)}{\partial \Delta x_i} = \lambda \frac{\partial g(x, \Delta x)}{\partial \Delta x_i}, \quad (\text{S16})$$

with a scalar Lagrange multiplier λ . Inserting Eqs. (S13) and (S15) and solving for the updates gives

$$\Delta x_i = \frac{1}{2\lambda a_i(x)^2} \frac{\partial F(x)}{\partial x_i}. \quad (\text{S17})$$

As we aim to maximize F , the updates must have a positive scalar product with the gradient of F . Thus, since the $a_i^2(x)$ are positive, λ must be positive. The condition that the updates satisfy the constraint Eq. (S14) then fixes λ : Inserting Eq. (S17) into Eq. (S14) gives

$$\lambda^2 = \frac{1}{4c(x)^2} \sum_{i=1}^N \frac{1}{a_i(x)^2} \left(\frac{\partial F(x)}{\partial x_i} \right)^2, \quad (\text{S18})$$

$$\lambda_{1,2} = \pm \frac{1}{2c(x)} \sqrt{\sum_{i=1}^N \frac{1}{a_i(x)^2} \left(\frac{\partial F(x)}{\partial x_i} \right)^2}, \quad (\text{S19})$$

since $c(x) > 0$. Inserting the positive solution into Eq. (S17) yields the updates

$$\Delta x_i = \frac{c(x)}{a_i(x)} \frac{\frac{1}{a_i(x)} \frac{\partial F(x)}{\partial x_i}}{\sqrt{\sum_{i=1}^N \frac{1}{a_i(x)^2} \left(\frac{\partial F(x)}{\partial x_i} \right)^2}}. \quad (\text{S20})$$

The formula shows that the updates are in the direction of the normalized gradient of F correcting for the costs for the updates. In particular, if the costs $a_i(x)$ for an update of the variable i are high, the update is reduced compared to a gradient ascent along F . Further, large gradients do not imply large update steps. Rather, the update step size is specified by the updating capabilities, c , and the costs

S5. DYNAMICAL EQUATIONS OF THE GREEDY DYNAMICS

The changes $\Delta_{g,H}a(t), \Delta_{g,H}R(t)$ that humankind applies to $a(t), R(t)$ during greedy dynamics maximize

$$F(a(t), R(t)) = a(t)R(t). \quad (\text{S21})$$

under the constraint

$$(\rho(R(t))\Delta_{g,H}R(t))^2 + (\alpha_H\Delta_{g,H}a(t))^2 = H(a(t), R(t))^2\Delta t^2. \quad (\text{S22})$$

More precisely speaking, the updates are such that they satisfy the constraint Eq. (S22) and maximize the (updated) objective $F(a(t) + \Delta_{g,H}a(t), R(t) + \Delta_{g,H}R(t))$, given that just before the update at time t the parameters were $a(t), R(t)$. Eq. (S20) thus yields the sizes of the updates when inserting $c(x) = H(a(t), R(t))\Delta t$, $N = 2$, $a_1(x) = \alpha_H$, $a_2(x) = \rho(R(t))$, $x_1 = a(t)$, $x_2 = R(t)$ and Eq. (S21):

$$\Delta_{g,H}a(t) = \frac{H(a(t), R(t))}{\alpha_H} \frac{\frac{1}{\rho(R(t))}R(t)}{n_{g,H}} \Delta t, \quad (\text{S23})$$

$$\Delta_{g,H}R(t) = \frac{H(a(t), R(t))}{\rho(R(t))} \frac{\frac{1}{\alpha_H}a(t)}{n_{g,H}} \Delta t, \quad (\text{S24})$$

with the normalization factor

$$n_{g,H} = \sqrt{\frac{1}{\alpha_H^2}R(t)^2 + \frac{1}{\rho(R(t))^2}a(t)^2}. \quad (\text{S25})$$

Analogously, the sizes of the updates $\Delta_{g,A}a(t)$ and $\Delta_{g,A}R(t)$ that AI applies are

$$\Delta_{g,A}a(t) = -\frac{A(a(t), R(t))}{\alpha_A} \frac{\frac{1}{\rho(R(t))}R(t)}{n_{g,A}} \Delta t, \quad (\text{S26})$$

$$\Delta_{g,A}R(t) = \frac{A(a(t), R(t))}{\rho(R(t))} \frac{\frac{1}{\alpha_A}(1 - a(t))}{n_{g,A}} \Delta t, \quad (\text{S27})$$

with the normalization factor

$$n_{g,A} = \sqrt{\frac{1}{\alpha_A^2}R(t)^2 + \frac{1}{\rho(R(t))^2}(1 - a(t))^2}. \quad (\text{S28})$$

S6. DYNAMICAL EQUATIONS OF THE FARSIGHTED MODEL

The changes $\Delta_H a(t), \Delta_H R(t)$ that humankind applies to $a(t), R(t)$ maximize

$$F(a(t), R(t)) = a(t + T_h)R(t + T_h). \quad (\text{S29})$$

under the constraint

$$(\rho(R(t))\Delta_H R(t))^2 + (\alpha_H \Delta_H a(t))^2 = H(a(t), R(t))^2 \Delta t^2. \quad (\text{S30})$$

More precisely speaking, the updates are such that they satisfy the constraint Eq. (S30) and maximize the (updated) expected future resources $F(a(t) + \alpha_H \Delta_H a(t), R(t) + \Delta_H R(t))$, given that just before the update at time t the state was $a(t), R(t)$. Eq. (S20) thus yields the sizes of the updates when inserting $c(x) = H(a(t), R(t))\Delta t$, $N = 2$, $a_1(x) = \alpha_H$, $a_2(x) = \rho(R(t))$, $x_1 = a(t)$, $x_2 = R(t)$ and Eq. (S29):

$$\Delta_H a(t) = \frac{H(a(t), R(t))}{\alpha_H} \frac{\frac{1}{\alpha_H} \frac{\partial a(t+T_h)R(t+T_h)}{\partial a(t)}}{n_H} \Delta t, \quad (\text{S31})$$

$$\Delta_H R(t) = \frac{H(a(t), R(t))}{\rho(R(t))} \frac{\frac{1}{\rho(R(t))} \frac{\partial a(t+T_h)R(t+T_h)}{\partial R(t)}}{n_H} \Delta t, \quad (\text{S32})$$

with the normalization factor

$$n_H = \sqrt{\frac{1}{\alpha_H^2} \left(\frac{\partial a(t+T_h)R(t+T_h)}{\partial a(t)} \right)^2 + \frac{1}{\rho(R(t))^2} \left(\frac{\partial a(t+T_h)R(t+T_h)}{\partial R(t)} \right)^2}. \quad (\text{S33})$$

Analogously, the sizes of the updates $\Delta_A a(t)$ and $\Delta_A R(t)$ that AI applies are

$$\Delta_A a(t) = \frac{A(a(t), R(t))}{\alpha_A} \frac{\frac{1}{\alpha_A} \frac{\partial (1-a(t+T_h))R(t+T_h)}{\partial a(t)}}{n_A} \Delta t, \quad (\text{S34})$$

$$\Delta_A R(t) = \frac{A(a(t), R(t))}{\rho(R(t))} \frac{\frac{1}{\rho(R(t))} \frac{\partial (1-a(t+T_h))R(t+T_h)}{\partial R(t)}}{n_A} \Delta t, \quad (\text{S35})$$

with the normalization factor

$$n_A = \sqrt{\frac{1}{\alpha_A^2} \left(\frac{\partial (1-a(t+T_h))R(t+T_h)}{\partial a(t)} \right)^2 + \frac{1}{\rho(R(t))^2} \left(\frac{\partial (1-a(t+T_h))R(t+T_h)}{\partial R(t)} \right)^2}. \quad (\text{S36})$$

For $T_h = 0$ these updates reduce to the ones of the greedy dynamics.

We obtain the derivatives of the future objectives through numerical differentiation. Specifically, we evolve the greedy dynamics from t to $t + T_h$ with perturbed initial conditions (I) $(a(t) + \varepsilon_a, R(t))$ to obtain $(a_I(t + T_h), R_I(t + T_h))$, (II) $(a(t) - \varepsilon_a, R(t))$ to obtain

$(a_{\text{II}}(t + T_h), R_{\text{II}}(t + T_h))$, (III) $(a(t), R(t) + \varepsilon_R)$ to obtain $(a_{\text{III}}(t + T_h), R_{\text{III}}(t + T_h))$ and (IV) $(a(t), R(t) - \varepsilon_R)$ to obtain $(a_{\text{IV}}(t + T_h), R_{\text{IV}}(t + T_h))$. Then the derivatives of the objectives of humankind and AI are

$$\frac{\partial a(t + T_h) R(t + T_h)}{\partial a(t)} \approx \frac{a_{\text{I}}(t + T_h) R_{\text{I}}(t + T_h) - a_{\text{II}}(t + T_h) R_{\text{II}}(t + T_h)}{2\varepsilon_a}, \quad (\text{S37})$$

$$\frac{\partial a(t + T_h) R(t + T_h)}{\partial R(t)} \approx \frac{a_{\text{III}}(t + T_h) R_{\text{III}}(t + T_h) - a_{\text{IV}}(t + T_h) R_{\text{IV}}(t + T_h)}{2\varepsilon_R}, \quad (\text{S38})$$

$$\frac{\partial (1 - a(t + T_h)) R(t + T_h)}{\partial a(t)} \approx \frac{(1 - a_{\text{I}}(t + T_h)) R_{\text{I}}(t + T_h) - (1 - a_{\text{II}}(t + T_h)) R_{\text{II}}(t + T_h)}{2\varepsilon_a}, \quad (\text{S39})$$

$$\frac{\partial (1 - a(t + T_h)) R(t + T_h)}{\partial R(t)} \approx \frac{(1 - a_{\text{III}}(t + T_h)) R_{\text{III}}(t + T_h) - (1 - a_{\text{IV}}(t + T_h)) R_{\text{IV}}(t + T_h)}{2\varepsilon_R}. \quad (\text{S40})$$

S7. MATCHING THE DYNAMICS TO EMPIRICAL DATA

We employ as measure for the produced electricity R available for humankind and AI the worldwide electricity consumption reduced by the electricity required for production, in the years 2024 (27391 TWh) and 2025 (28199 TWh) [3, pp. 15, 216]. These values yield in 2024 and 2025 average monthly available amounts of $\bar{R}_{2024} \approx 8.22$ EJ/month and $\bar{R}_{2025} \approx 8.46$ EJ/month, respectively, which are highlighted by the horizontal dashed lines in Fig. S7c and Fig. S10c. To obtain an approximal initial time course, we assume an exponential increase of the available electricity R over these two years, setting

$$R_{\text{init}}(t) = R_{\text{init},0} e^{\lambda_{\text{init}} t}. \quad (\text{S41})$$

We compute the values of $R_{\text{init},0}$ and λ_{init} from the conditions that the averages of $R_{\text{init}}(t)$ over the first and second year match $\bar{R}_{2024}$ and $\bar{R}_{2025}$, respectively. The resulting $R_{\text{init}}(t)$, with

$$R_{\text{init},0} \approx 8.10 \text{ EJ/month and } \lambda_{\text{init}} \approx 0.029/\text{year}, \quad (\text{S42})$$

is displayed by dashed curves in Fig. S7c and Fig. S10c.

The average energy consumption of AI is difficult to assess and estimates vary greatly. We adopt an intermediate value of an overall consumption of 100 TWh in 2024 [9, p. 57], and a yearly growth by 50%, which was observed in 2025 [6, p. 9]. This yields average monthly consumptions of $\bar{R}_{A,2024} = 0.030$ EJ/month and $\bar{R}_{A,2025} = 0.045$ EJ/month. We transform them into estimates of a using $R_A = (1 - a)R$, which yields $\bar{a}_{2024} \approx 1 - \bar{R}_{A,2024}/\bar{R}_{2024} \approx 0.9963$ and $\bar{a}_{2025} \approx 0.9947$; these are highlighted by dashed horizontal lines in Fig. S7d and Fig. S10d. To obtain an initial expected time course of a , we compute an initial time course of R_A as for R , i.e. we set (see also Eq. (S4))

$$R_{A,\text{init}}(t) = R_{A,\text{init},0} e^{\lambda_{A,\text{init}} t}, \quad (\text{S43})$$

where

$$R_{A,\text{init},0} \approx 0.024 \text{ EJ/month and } \lambda_{A,\text{init}} = \ln(1.5)/\text{year}, \quad (\text{S44})$$

such that the averages $\bar{R}_{A,2024}$ and $\bar{R}_{A,2025}$ are matched. This gives

$$a_{\text{init}}(t) = 1 - \frac{R_{A,\text{init}}(t)}{R_{\text{init}}(t)}, \quad (\text{S45})$$

which is displayed by dashed curves in Fig. S7d and Fig. S10d. If not mentioned otherwise, we use

$$R_0 = R_{\text{init}}(0) \approx 8.10 \text{ EJ/month}, \quad (\text{S46})$$

$$a_0 = a_{\text{init}}(0) \approx 0.9970 \quad (\text{S47})$$

as initialization of $R(t)$ and $a(t)$, setting $R(0) = R_0$ and $a(0) = a_0$.

We choose the model parameters such that the resulting initial efficacy of humanity $H(a(0)R(0))$ is considerably larger than the efficacy $A(a(0)R(0))$ of AI, which is plausible in the beginning of 2024, where our simulations start. Further, we assume that the cost of modifying the distribution of produced resources is much larger for AI than for humanity, as humanity controls resource distribution in the beginning and AI needs to take control. Beyond this we choose the parameters α_H , α_A , c_H and c_A such that the time course of the full model dynamics matches over the first two years that of the empirically observed available electricity, $R_{\text{init}}(t)$, and the fraction of electricity obtained by humankind without AI, $a_{\text{init}}(t)$, in 2024 and 2025. Concretely we set $\alpha_H = 70$ and $\alpha_A = 500$ and c_H and c_A such that $H(a(0)R(0)) = 0.02$ and $A(a(0)R(0)) = 0.0059$.

S8. ILLUSTRATION OF THE GREEDY DYNAMICS

The section provides illustrations of the greedy dynamics. Fig. S3 shows the ensuing greedy dynamics for our standard initial conditions: if humankind and AI were optimizing their resources in greedy manner, humankind would eliminate AI. This is because in order to maximize $a(t + \Delta t)R(t + \Delta t)$ for small Δt , humankind generates positive updates $\Delta_H a(t)$ and $\Delta_H R(t)$ to $a(t)$ and $R(t)$. AI counteracts by decreasing $a(t)$ and increasing $R(t)$ to maximize $(1 - a(t + \Delta t))R(t + \Delta t)$, but due to the larger initial efficacy of humankind and, in addition, lower redistribution costs compared to AI, $a(t)$ increases to 1. If the initial share of AI was considerably larger, AI would marginalize humankind by decreasing $a(t)$ to zero, see Fig. S4.

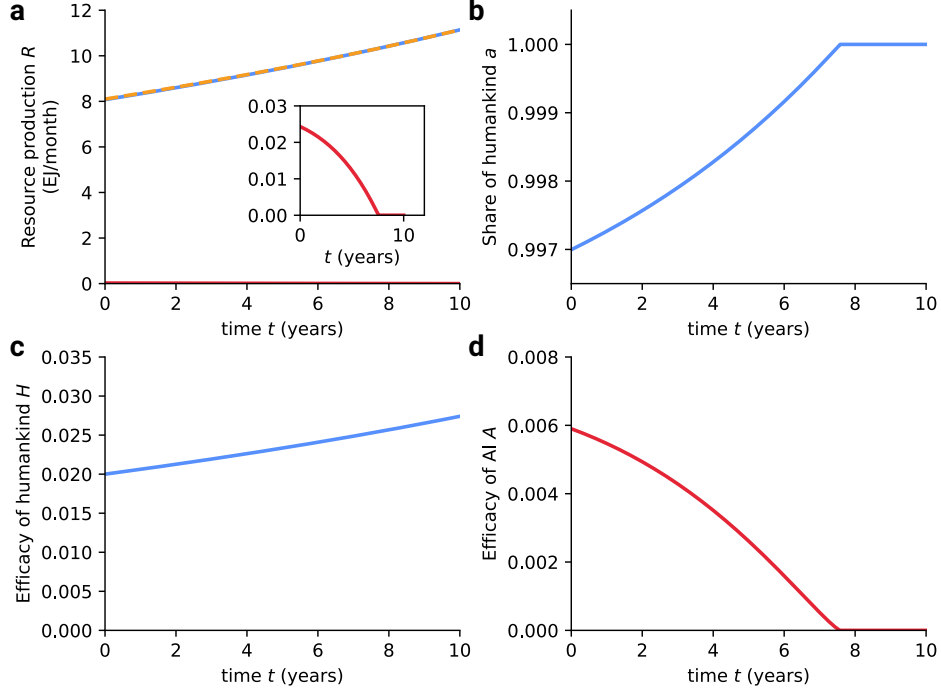


Figure S3. Greedy dynamics. (a) When assuming greedy dynamics, the total amount of produced resources R steadily increases (orange dashed curve). The vast majority of produced resources go to humankind (part aR , blue curve, overlaid by orange curve), which dominates the dynamics and mainly drives the increase of R . Remaining resources go to AI (part $(1-a)R$, red curve), magnified in Inset. This part is steadily reduced and reaches zero, i.e. humankind eliminates AI to use the resources itself. (b) Humankind increases its share a of produced resources against the counteraction of AI, which results in the elimination of AI as a reaches one. (c) Efficacy H of humankind steadily increases. (d) AI's efficacy A decreases to zero as it is depleted of resources.

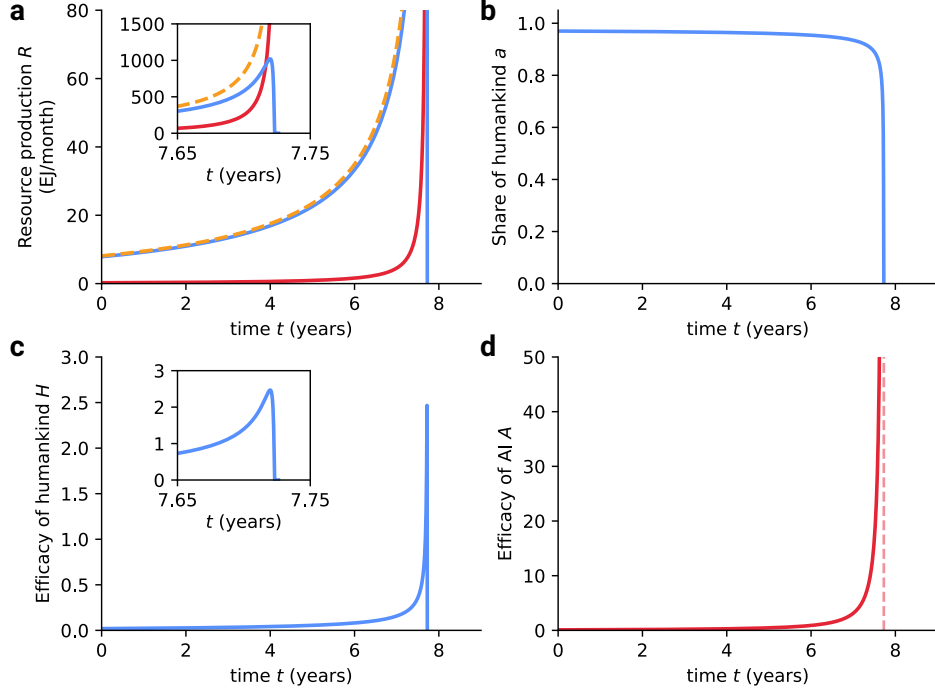


Figure S4. Greedy dynamics with reduced initial share of humankind's resources a , such that AI initially dominates. (a) R (orange dashed) increases quickly; it tends to infinity in finite time (singularity). The part of resources received by humankind (aR , blue, partially overlaid by orange curve) initially increases strongly as well. Shortly before the singularity, it falls to zero, i.e. AI marginalizes humanity. The part received by AI increases steeply ($(1-a)R$, red). (b) a decreases initially slowly then steeply, to zero. (c) Efficacy H of humankind increases transiently much more strongly than for dominating humankind, compare Fig. S3c. This is because of the strong increase of produced resources obtained by AI, which leads to an increase of aR despite the simultaneous reduction of a . As a finally tends to zero, the efficacy of humankind, however, falls to a small baseline value. The remaining part is because some efficacy remains also without the resource of electricity, which we are focusing on, compare Fig. S1 and Eq. (S1). (h) The efficacy of AI strongly increases; it tends to infinity at a finite time point (red dashed vertical line).

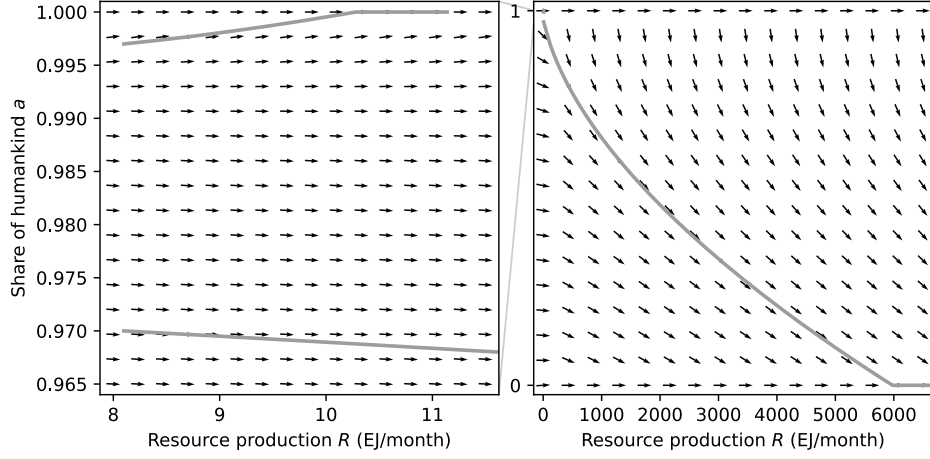


Figure S5. State space plot of the greedy dynamics. The plot displays $R(t)$ and $a(t)$ of Fig. S3 and of Fig. S4 as upper and lower parametric curves in R - a state space, respectively. Directions of change when hypothetically passing a point, i.e. directions of velocity vector $(dR(t)/dt, da(t)/dt)$, are shown as arrows at selected points. They indicate the basins of attraction of elimination of AI and marginalization of mankind. Left hand side displays a closeup of the region with a close to one and R close to R_0 .

S9. FURTHER DETAIL ON THE DYNAMICS OF MAIN TEXT FIG. 1

The section provides further illustrations of the full dynamics with our standard planning horizon of $T_h = 3$ year. Fig. S6 shows the trajectories of the dynamics within state space and as vector field the direction of changes per time. Fig. S7 dissects the changes in a and R into contributions of humankind and AI. Furthermore, it illustrates the match of the dynamics to empirical data, as described in Sec. S7. Fig. S8 shows the changes of humankind's future resources with $a(0)$, clarifying why humankind initially fosters AI in the full dynamics in contrast to the greedy dynamics.

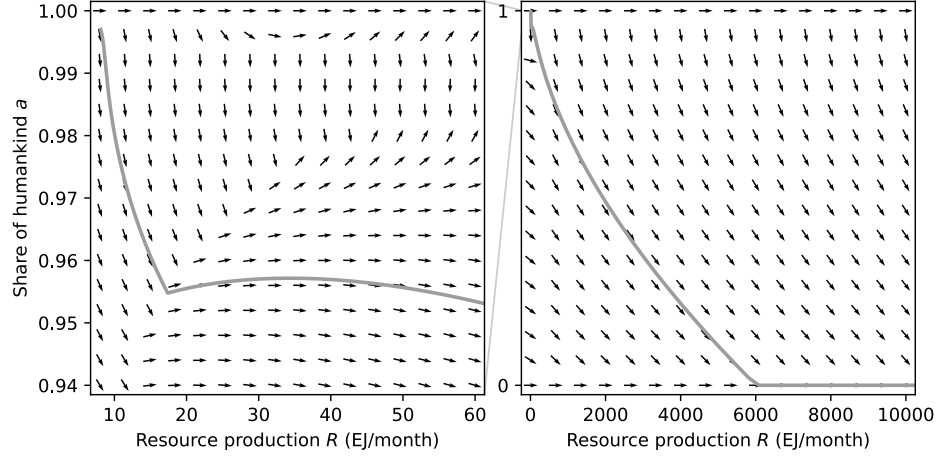


Figure S6. State space plot of the dynamics of Fig. 1. The plot displays $R(t)$ and $a(t)$ as parametric curves in R - a state space. The direction of changes, i.e. the direction of the velocity vectors $(dR(t)/dt, da(t)/dt)$, are displayed by arrows. Left hand side displays a closeup of the region with a close to one and smaller R .

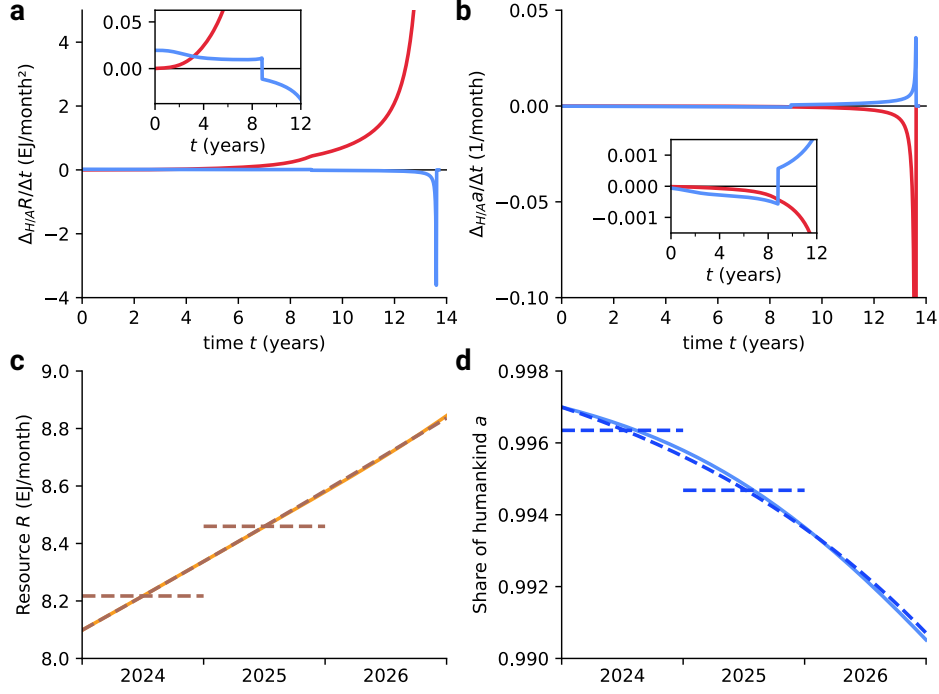


Figure S7. Impact of humankind and AI on the dynamics of main text Fig. 1 and matching of the initial dynamics to empirical data. (a) Contribution $\Delta_H R$ (blue, Eq. (S32)) and $\Delta_A R$ (red, Eq. (S35)) of humankind and AI, respectively, to the total change ΔR of R , divided by update step size Δt , as a function of time. For early times, the contribution of humankind dominates, but it is greatly exceeded by the contribution of AI towards the end. At some point, humankind starts to decrease R (Inset), since this delays marginalization, which is advanced by AI on the basis of large R . Close to marginalization, humankind's contribution quickly drops to zero. AI increases R , first by small, later, when its efficacy has increased, by large amounts. (b) Contribution $\Delta_H a$ (blue, Eq. (S31)) and $\Delta_A a$ (red, Eq. (S34)) of humankind and AI, respectively, to the total change Δa of a , divided by update step size Δt , as a function of time. At early times, both humankind and AI decrease a (Inset). Humankind does this because it foresees that contributing resources to AI now will lead to larger resources available to humankind in the future. This is because AI uses these resources better than humankind would. More technically speaking, humankind's predicted resources $a(t + T_h)R(t + T_h)$ increase when it decreases the current $a(t)$, see also Fig. S8. (c) Monthly generated resource R (continuous curve) over the first three years, compared to empirically measured averages of the worldwide available electricity production in 2024 and 2025 (dashed horizontal lines, from [3]) and exponential interpolation of averages (dashed curve). (d) Fraction a of generated resource R received by humankind. The panel shows the time course of a in the model (continuous curve) together with empirically estimated averages in 2024 and 2025 (dashed horizontal lines, from [6, 9]) and an interpolation of these averages assuming an exponential increase of both the resources obtained by humanity and AI.

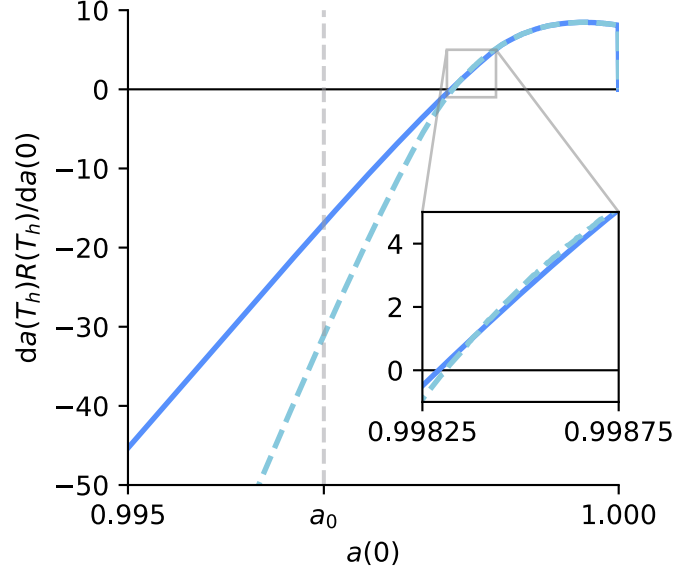


Figure S8. Change of humankind's projected future resources when changing the initial resource allocation, as a function of the initial fraction of humankind's resources $a(0)$. At the actual initial values of a and R , $a(0) = a_0$ (dashed vertical line) and $R(0) = R_0$, the projected resources at T_h decrease if $a(0)$ slightly increases: the derivative $da(T_h)R(T_h)/da(0)$ is negative. In other words, it is more beneficial for humankind to allocate some of its resources at $t = 0$ to AI, as AI generates more resources from this at T_h than humankind could. This outweighs the loss in humankind's future resources $a(T_h)R(T_h)$ due to the reduced initial $a(0)$. As a consequence, humankind reduces a during the initial phase of the dynamics, fostering AI. Thus, in contrast to the greedy dynamics, the initially dominating humankind increases the resources that AI receives. This holds for both considered planning horizons $T_{h,0} = 3\text{year}$ (blue) and $T_{h,0} = 5\text{year}$ (light blue dashed). For larger $a(0)$, it is more beneficial for humankind to keep its resources, because AI is too weak to generate at due time, i.e. at T_h , more resources than humankind would: $da(T_h)R(T_h)/da(0)$ is positive. Inset: The dashed curve crosses zero at slightly larger $a(0)$, since the period of planning ahead and thus the time available for AI to become useful are longer. (The initial value of R is $R(0) = R_0$ throughout.)

S10. FURTHER DETAIL ON THE DYNAMICS OF MAIN TEXT FIG. 2

The section provides further illustrations of the full dynamics with extended planning horizon $T_h = 5$ year. Analogous to Fig. S7, Fig. S10 shows the contributions of humankind and AI to the changes in a and R and how the initial dynamics fit empirical data.

In Fig. S10a and Fig. S10b humankind's updates of R and a partially oscillate irregularly between positive and negative values. These oscillations originate from the fact that humankind's resources in the greedy dynamics with AI domination have a peak towards the end, see Fig. S4a Inset: If humankind's planning phase ends at the rising part of the peak, it appears beneficial to foster AI, because this moves the peak to earlier times and thus seems to increase the resources at the planning phase's end. If the planning phase is behind the peak, it analogously appears beneficial to hinder AI, such that the peak moves to later times. Generating resources R and decreasing a fosters AI, reducing resources and increasing a hinders it. In a period where humankind fosters AI, the peak moves earlier until the planning phase's end falls behind the peak, such that it appears beneficial to hinder AI. Humankind then hinders AI, until the peak moves so far to later times that the planning phase's end falls before the peak, and it appears again beneficial to foster AI.

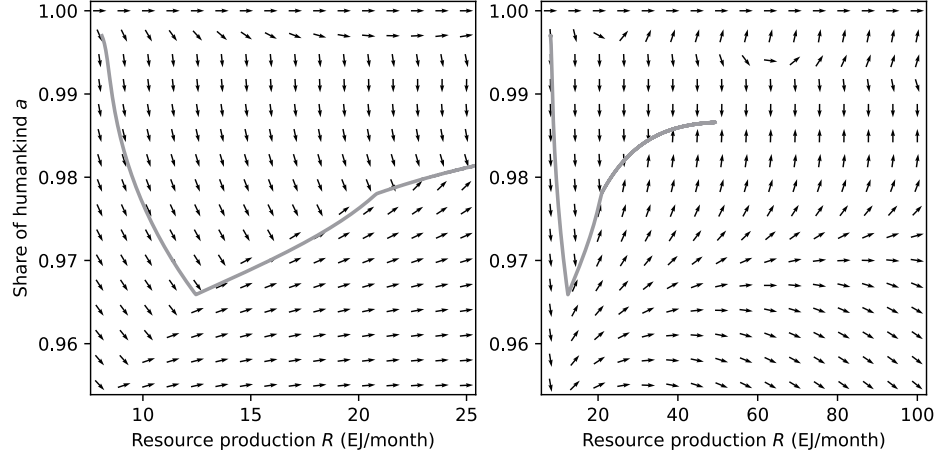


Figure S9. State space plot of the dynamics of Fig. 2. The plot displays $R(t)$ and $a(t)$ as parametric curves in the R - a state space. The direction of changes, i.e. the direction of the velocity vectors $(dR(t)/dt, da(t)/dt)$, are displayed by arrows. They indicate the basin of attraction of the state of balanced growth. Both plots show a region with a close to one, the left hand side displays a region with R closer to R_0 .

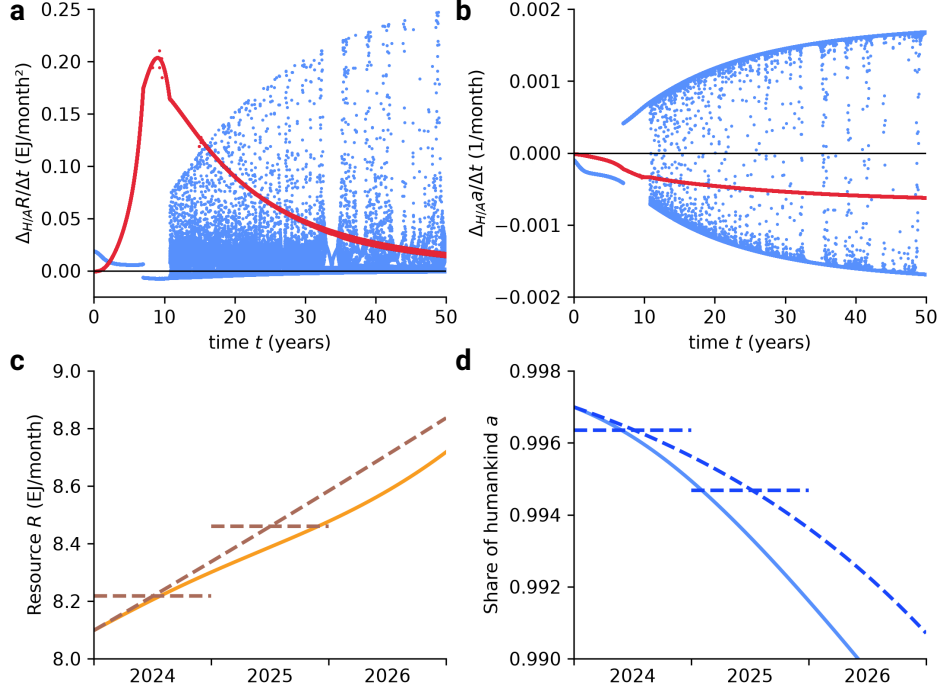


Figure S10. Impact of humankind and AI on the dynamics of main text Fig. 2 and matching of the initial dynamics to empirical data. (a) Panel analogous to Fig. S7a, but with different time and update scales. For early times, the contribution of humankind dominates, but it is greatly exceeded by the contribution of AI at intermediate times. AI increases R , while humankind first increases R , then decreases it, to prevent overly fast AI growth and marginalization. Finally humankind's updates oscillate between increase and decrease, because it sometimes appears beneficial to foster and sometimes to hinder AI. b) Panel analogous to Fig. S7b, but with different time and update scales. In the beginning, humankind decreases a , fostering AI. Analogous to (a) humankind thereafter increases a , to prevent overly fast AI growth and marginalization. Finally, its updates oscillate between decrease and increase. AI decreases a throughout. c),d) Panels analogous to Fig. S7c,d. The match of R and a (continuous) to the empirical estimates (dashed) is worse compared to Fig. S7c,d, because T_h has been increased without adapting further parameters. In particular, a decreases more steeply, since humankind expects more future benefit of AI, as AI's efficacy has more time to grow. This stronger projection into the future leads to a reduction of the increase of actually produced resources (continuous orange curve in (a) is lower than dashed brown curve).

S11. MODEL PARAMETERS

Parameter	Value	Unit
$c_{\text{el. avail.}}$	0.8855737574534686	
m	11.300198969875929	$10^{12}\text{USD} \times \text{month/EJ}$
b	2.3084355749119707	10^{12}USD
l_0	28.5	year
m_l	26.033595689869003	$\text{year} \times (\text{month/EJ})^{\kappa_l}$
κ_l	0.26865811057100736	
$\bar{R}_{2024}$	8.2173	EJ/month
$\bar{R}_{2025}$	8.4597	EJ/month
$\bar{R}_{A,2024}$	0.03	EJ/month
$\bar{R}_{A,2025}$	0.045	EJ/month
$R_0 = R(0) = R_{\text{init},0} = R_{\text{init}}(0)$	8.098431980182847	EJ/month
$R_{A,\text{init},0} = R_{A,\text{init}}(0)$	0.02432790648648986	EJ/month
λ_{init}	0.02907202403881548	1/year
$\lambda_{A,\text{init}}$	$\ln(1.5)$	1/year
$\bar{a}_{2024}$	0.9963491657843817	
$\bar{a}_{2025}$	0.9946806624348381	
$a_0 = a(0) = a_{\text{init},0} = a_{\text{init}}(0)$	0.9969959732271604	
α_H	70	
α_A	500	
c_H	0.00021379531796626502 ($H(a(0)R(0)) = 0.02$)	$1/(10^{12}\text{USD})$
c_A	0.5099485090045803 ($A(a(0)R(0)) = 0.0059$)	
T_h	3 or 5	year
κ	1.2	
Δt	$12/(365 \times 24)$	month

Table S1. The table displays the values of the model parameters, in machine precision as used in the figures and simulations.

-
- [1] R. Pinto, S. T. Henriques, P. E. Brockway, M. K. Heun, and T. Sousa, The rise and stall of world electricity efficiency:1900–2017, results and insights for the renewables transition, Energy

269, 126775 (2023).

- [2] Ember, Yearly electricity data, <https://ember-energy.org/data/yearly-electricity-data/> (2026), accessed: Mar 18, 2026.
- [3] International Energy Agency, *Electricity 2026: Analysis and forecast to 2030*, Tech. Rep. (International Energy Agency (IEA), Paris, France, 2026).
- [4] S. Dattani, L. Rodés-Guirao, H. Ritchie, E. Ortiz-Ospina, and M. Roser, Economic growth: Gross domestic product (gdp), Our World in Data (2023), <https://archive.ourworldindata.org/20260518-093348/grapher/gdp-worldbank-constant-usd.html>.
- [5] T. Kwa, B. West, J. Becker, A. Deng, K. Garcia, M. Hasin, S. Jawhar, M. Kinniment, N. Rush, S. Von Arx, R. Bloom, T. Broadley, H. Du, B. Goodrich, N. Jurkovic, L. H. Miles, S. Nix, T. Lin, N. Parikh, D. Rein, L. J. K. Sato, H. Wijk, D. M. Ziegler, E. Barnes, and L. Chan, Measuring AI ability to complete long software tasks (2025), arXiv:2503.14499 [cs.AI].
- [6] International Energy Agency, *Key Questions on Energy and AI*, Tech. Rep. (International Energy Agency (IEA), Paris, France, 2026).
- [7] S. Dattani, L. Rodés-Guirao, H. Ritchie, E. Ortiz-Ospina, and M. Roser, Life expectancy, Our World in Data (2023), <https://ourworldindata.org/life-expectancy>.
- [8] M. Faber and E. Virguez-Rodriguez, Supporting decisions on global health and life safety investments, in *Applications of Statistics and Probability in Civil Engineering* (CRC Press, 2011) pp. 434–443.
- [9] International Energy Agency, *Energy and AI*, Tech. Rep. (International Energy Agency (IEA), Paris, France, 2025).